

Selection of variables and decision boundaries for functional data via bi-level selection

Hidetoshi Matsui

*The Center for Data Science Education and Research, Shiga University
1-1-1, Banba, Hikone, Shiga 522-8522, Japan.*

hmatsui@biwako.shiga-u.ac.jp

Abstract: Sparsity-inducing penalties are useful tools for variable selection and they are also effective for regression settings where the data are functions. We consider the problem of selecting not only variables but also decision boundaries in logistic regression models for functional data, using the sparse regularization. The functional logistic regression model is estimated by the framework of the penalized likelihood method with the sparse group lasso-type penalty, and then tuning parameters are selected using the model selection criterion. The effectiveness of the proposed method is investigated through real data analysis.

Key Words and Phrases: Functional data analysis, Model selection, Sparse group lasso

1 Introduction

Variable selection is one of the most important issues in regression analysis and several methods have been proposed for the accurate and effective selection of appropriate variables (see, e.g., Konishi and Kitagawa, 2008). For such problems, sparse regularization that estimates the model with L_1 -type penalties provides a unified approach for estimating and selecting variables, and for this reason they are broadly applied in several fields (Bühlmann and van de Geer, 2011; Hastie et al., 2015). In this paper, we consider the problem of selecting not only variables but also decision boundaries which affect the classification problem, by applying the sparse regularization to logistic regression models when the data to be classified are measured repeatedly over time.

The logistic regression model is a useful tool for classifying data, and it does so by providing posterior probabilities which place the data in the appropriate group (McCullagh and Nelder, 1989). Logistic regression modeling that use the sparse regularization have been investigated as generalized linear models in Park and Hastie (2007) and Krishnapuram et al. (2005), and Friedman et al. (2010b) applied L_1 -type penalties to the multinomial or multiclass logistic regression model that classifies data into three or more groups as natural extensions of the binomial logistic regression models. More recently, Vincent and Hansen (2014) applied the sparse group lasso-type penalty (Simon et al., 2013b) to the logistic regression model. The sparse group lasso is one of the bi-level selection techniques

(Breheny and Huang, 2009; Matsui, 2015) that select variables in both group and individual levels. Therefore, it can be seen as a composition of the lasso (Tibshirani, 1996) and the group lasso (Yuan and Lin, 2006).

Functional data analysis (FDA), which is established by Ramsay and Silverman (2005), is a useful method for effectively analyzing repeatedly measured data, and it has received considerable attentions in various fields (Ramsay and Silverman, 2002; Horváth and Kokoszka, 2012). The basic idea behind FDA is to express repeated measurement data for each individual as a smooth function and then to draw information from the collection of these functions. FDA includes extensions of traditional methods, and in particular there are many works on regression models. For logistic regression models for functional data, there are various works in Aguilera and Escabias (2008), Aguilera-Morillo et al. (2013), and Escabias et al. (2007). Furthermore, the problem of variable selection for functional regression models using L_1 -type regularization is considered in Ferraty et al. (2010), Aneiros et al. (2011), Matsui and Konishi (2011), Zhao et al. (2012), Gertheiss et al. (2013), and Mingotti et al. (2013). However, these works do not include the multiclass logistic regression model. For this model, we may fail to select functional variables when we use existing types of penalties, since it has multiple coefficients for multiple decision boundaries. In order to solve this problem, Matsui (2014) proposed two types of penalties for selecting variables and decision boundaries respectively.

In this paper we apply the bi-level selection technique to the functional logistic regression model in order to select variables and decision boundaries simultaneously. Time course observations are smoothed by using basis expansions, and then parameters included in the functional logistic regression model are estimated by the sparse regularization with the sparse group lasso-type penalty. We apply the blockwise descent algorithm derived by Simon et al. (2013b) for estimating the coefficient parameters. Values of tuning parameters in the penalty function are selected by a model selection criterion. The effectiveness of the proposed method is investigated through the real data analysis.

This paper is organized as follows. In Section 2, we introduce the details of the logistic regression model for functional data and some preparations for estimation of the model. Section 3 provides the method for estimating coefficient parameters and for selecting tuning parameters. The results of real data analysis are discussed in Section 4. Finally we conclude the article with some discussions in Section 5.

2 Multiclass logistic regression model for functional data

Suppose we have n sets of functional data and class labels $\{(x_i(t), g_i); i = 1, \dots, n\}$, where $x_i(t) = (x_{i1}(t), \dots, x_{ip}(t))^T$ are predictors given as functions and $g_i \in \{1, \dots, L\}$ are classes to which each x_i belongs. In the classification setting, we apply the Bayes rule, which assigns x_i to class $g_i = l$ with the maximum posterior probability given x_i , denoted by $\Pr(g_i = l|x_i)$. Then the functional logistic regression model is given by the log-odds of the posterior probabilities:

$$\log \left\{ \frac{\Pr(g_i = l|x_i)}{\Pr(g_i = L|x_i)} \right\} = \beta_{0l} + \sum_{j=1}^p \int x_{ij}(t) \beta_{jl}(t) dt, \quad (1)$$

where β_{0l} is an intercept and $\beta_{jl}(t)$ are coefficient functions. We assume that $x_{ij}(t)$ can be expressed by basis expansions as

$$x_{ij}(t) = \sum_{m=1}^{M_j} w_{ijm} \phi_{jm}(t) = w_{ij}^T \phi_j(t), \quad (2)$$

where $\phi_j(t) = (\phi_{j1}(t), \dots, \phi_{jM_j}(t))^T$ are vectors of basis functions such as B -splines or radial basis functions, and $w_{ij} = (w_{ij1}, \dots, w_{ijM_j})^T$ are coefficient vectors. Since the data are originally observed at discrete time points, we smooth them with a basis expansion prior to obtaining the functional data $x_{ij}(t)$. In other words, w_{ij} are obtained before constructing the functional logistic regression model (1). Details of the smoothing method are described in Araki et al. (2009b). Furthermore, $\beta_{jl}(t)$ are also expressed by basis expansions

$$\beta_{jl}(t) = \sum_{m=1}^{M_j} b_{jlm} \phi_{jm}(t) = b_{jl}^T \phi_j(t), \quad (3)$$

where $b_{jl} = (b_{jl1}, \dots, b_{jlM_j})^T$ are vectors of coefficient parameters.

Using the notation $\pi_l(x_i; b) = \Pr(g_i = l|x_i)$, where $b = (b_1^T, \dots, b_p^T)^T$ and $b_j = (b_{j1}^T, \dots, b_{j(L-1)}^T)^T$ since it is controlled by b , we can express the functional logistic regression model (1) as

$$\log \left\{ \frac{\pi_l(x_i; b)}{\pi_L(x_i; b)} \right\} = \beta_{0l} + \sum_{j=1}^p w_{ij}^T \Phi_j b_{jl} = \sum_{j=1}^p z_{ij}^T b_{jl}, \quad (4)$$

where $\Phi_j = \int \phi_j(t) \phi_j^T(t) dt$ and $z_{ij} = w_{ij}^T \Phi_j$. It follows from (1) that the posterior proba-

bility is

$$\pi_l(x_i; b) = \frac{\exp\left(\sum_j z_{ij}^T b_{jl}\right)}{1 + \sum_{h=1}^{L-1} \exp\left(\sum_j z_{ij}^T b_{jh}\right)} \quad (l = 1, \dots, L-1),$$

$$\pi_L(x_i; b) = \frac{1}{1 + \sum_{h=1}^{L-1} \exp\left(\sum_j z_{ij}^T b_{jh}\right)}.$$

We define the vectors of the response variables y_i , which indicate the class labels, as

$$y_i = (y_{i1}, \dots, y_{i(L-1)})^T = \begin{cases} (0, \dots, 0, \overset{(l)}{1}, 0, \dots, 0)^T & \text{if } g_i = l, \quad l = 1, \dots, L-1, \\ (0, \dots, 0)^T & \text{if } g_i = L. \end{cases}$$

Then the functional logistic regression model has the probability function

$$f(y_i|x_i; b) = \prod_{l=1}^{L-1} \pi_l(x_i; b)^{y_{il}} \pi_L(x_i; b)^{1 - \sum_{h=1}^{L-1} y_{ih}}. \quad (5)$$

3 Estimation by sparse regularization

From the result of the previous section we can construct a likelihood function $\ell(b) = \sum_i \log f(x_i; b)$. This can be expressed as

$$\ell(b) = -\frac{1}{2} \left\| W^{1/2} (\eta - \tilde{Z}b) \right\|_2^2, \quad (6)$$

where $W = (W_{hl})$ with

$$W_{hl} = \begin{cases} \text{diag} \{ \pi_l(x_1; b)(1 - \pi_l(x_1; b)), \dots, \pi_l(x_n; b)(1 - \pi_l(x_n; b)) \} & (h = l) \\ \text{diag} \{ -\pi_h(x_1; b)\pi_l(x_1; b), \dots, -\pi_h(x_n; b)\pi_l(x_n; b) \} & (h \neq l), \end{cases}$$

and $W^{1/2}$ is a matrix that satisfies $W = W^{1/2} W^{1/2}$. Furthermore, $\tilde{Z} = (\tilde{Z}_1, \dots, \tilde{Z}_p)$ with $\tilde{Z}_j = I_{L-1} \otimes Z_j$ and $Z_j = (z_{1j}, \dots, z_{nj})^T$, $\eta = \tilde{Z}b + W^{-1} \Lambda 1_{n(L-1)}$, $\Lambda = \text{diag} \{ \Lambda_1, \dots, \Lambda_{L-1} \}$, $\Lambda_l = \text{diag} \{ y_{1l} - \pi_l(x_1; b), \dots, y_{nl} - \pi_l(x_n; b) \}$, and $1_{n(L-1)} = (1, \dots, 1)^T$ is an $n(L-1)$ -dimensional vector. Then we consider maximizing the penalized log-likelihood function

$$\ell_{\lambda, \alpha}(b) = \ell(b) - P_{\lambda, \alpha}(b), \quad (7)$$

where we assume the sparse group lasso-type penalty for $P_{\lambda, \alpha}(b)$:

$$P_{\lambda, \alpha}(b) = n(1 - \alpha) \sum_{j=1}^p \lambda_j \left\{ \sum_{l=1}^{L-1} \|b_{jl}\|_2^2 \right\}^{1/2} - n\alpha \sum_{j=1}^p \lambda_j \sum_{l=1}^{L-1} \|b_{jl}\|_2, \quad (8)$$

where $\lambda_j = \sqrt{M_j} \lambda$ with a regularization parameter $\lambda > 0$ and $\alpha \in [0, 1]$ is a tuning parameter. The first term of this penalty has an effect that it shrinks some of b_j towards

exactly zero vectors, using the idea of the group lasso, and it leads to variable selection, while the second term shrinks b_{jl} toward zero vectors separately, which leads to decision boundary selection.

We want to estimate b by maximizing the function (7), but there are two difficulties in deriving the estimator. First, it is generally difficult to explicitly express the parameters estimated by the sparse regularization. In order to solve this problem we use the idea of the coordinate descent algorithm (Friedman et al., 2007). Second, when we apply the sparse group lasso-type penalty it is difficult to construct updated values for parameters if the design matrices for each of the groups (in our case $\tilde{Z}_1, \dots, \tilde{Z}_p$) are not orthogonal (Friedman et al., 2010a). Simon et al. (2013a) approached this problem by applying the Taylor expansion and the majorization-minimization algorithm. On the other hand, we apply the QR decomposition by using the idea of Simon and Tibshirani (2012) to form the orthogonal design matrix. The QR decomposition provides $W^{1/2}\tilde{Z}_j = Q_j R_j$, where Q_j is an orthogonal matrix and R_j is an upper triangle matrix. Denote $b_j^* = R_j b_j$, then the log-likelihood function (6) can be re-expressed by

$$\ell(b^*) = -\frac{1}{2} \left\| W_j^{1/2} r_{-j} - Q_j b_j^* \right\|_2^2,$$

where $r_{-j} = \eta - \sum_{j' \neq j} Z_{j'} b_{j'}$.

The partial derivative of $\ell_{\lambda, \alpha}(b^*)$ with respect to b_j^* is given by

$$\frac{\partial \ell_{\lambda}(b^*)}{\partial b_j^*} = \tilde{r}_{-j} - b_j^* - n(1 - \alpha)\lambda_j u_j - n\alpha\lambda_j v_j,$$

where $\tilde{r}_{-j} = (\tilde{r}_{-j1}, \dots, \tilde{r}_{-j(L-1)})^T = Q_j^T W_j^{1/2} r_{-j}$ and u_j and $v_j = (v_{j1}, \dots, v_{j(L-1)})^T$ are vectors of subgradients respectively given by

$$u_j = \begin{cases} \frac{b_j^*}{\|b_j^*\|_2} & (b_j^* \neq 0) \\ \text{s.t. } \|u_j\|_2 \geq 1 & (b_j^* = 0), \end{cases}$$

$$v_{jl} = \begin{cases} \frac{b_{jl}^*}{\|b_{jl}^*\|_2} & (b_{jl}^* \neq 0) \\ \text{s.t. } \|v_{jl}\|_2 \geq 1 & (b_{jl}^* = 0). \end{cases}$$

Let $S_j = (S_{j1}, \dots, S_{j(L-1)})^T$ with $S_{jl} = (\|\tilde{r}_{jl}\|_2 - n\alpha\lambda)_+$ be vectors of thresholding functions, where $(a)_+ = \max\{a, 0\}$, then if $\|S_j\|_2 \leq n(1 - \alpha)\lambda$ the parameter vector b_j^* is estimated to be $\hat{b}_j^* = 0$. Otherwise, solve the following equation with respect to b_{jl}^* :

$$\frac{\partial \ell_{\lambda}(b^*)}{\partial b_{jl}^*} = \tilde{r}_{-jl} - b_{jl}^* - n(1 - \alpha)\lambda_j \frac{b_{jl}^*}{\|b_j^*\|_2} - n\alpha\lambda_j v_{jl} = 0.$$

Then, if $\|\tilde{r}_{jl}\|_2 \leq n\alpha\lambda$ then $\hat{b}_{jl}^* = 0$, otherwise \hat{b}_{jl}^* is calculated as

$$\hat{b}_{jl}^* = \frac{\|\hat{b}_j^*\|_2(\|\tilde{r}_{-jl}\|_2 - n\alpha\lambda)_+}{\|\hat{b}_j^*\|_2 + n(1 - \alpha)\lambda} \frac{\tilde{r}_{-jl}}{\|\tilde{r}_{-jl}\|_2},$$

where $\|\hat{b}_j^*\|_2$ is given by

$$\|\hat{b}_j^*\|_2 = (\|h_j\| - n(1 - \alpha)\lambda)_+,$$

$$h_j = (h_{j1}^T, \dots, h_{j(L-1)}^T)^T, h_{jl} = (\|\tilde{r}_{-jl}\|_2 - n\alpha\lambda)_+ \frac{\tilde{r}_{-jl}}{\|\tilde{r}_{-jl}\|_2}.$$

The algorithm is given in the following steps:

1. (Outer loop) For $j = 1, \dots, p$, check if $\|S_j\|_2 \leq n(1 - \alpha)\lambda$. If it is true, $\hat{b}_j^* = 0$ and if not, for each j go to Step 2.
2. (Inner loop) For $l = 1, \dots, L - 1$, update b_{jl}^* as follows:

$$\hat{b}_{jl}^* = \frac{\|\hat{b}_j^*\|_2(\|\tilde{r}_{-jl}\|_2 - n\alpha\lambda)_+}{\|\hat{b}_j^*\|_2 + n(1 - \alpha)\lambda} \frac{\tilde{r}_{-jl}}{\|\tilde{r}_{-jl}\|_2}.$$

3. Iterate Step 1 and 2 until convergence and then obtain estimators $\hat{b}_1^*, \dots, \hat{b}_p^*$.
4. Calculate $\hat{b}_j = R_j^{-1}b_j^*$ for each j .

The outer loop corresponds to the variable selection step, and the inner loop corresponds to the decision boundary selection step.

The statistical model estimated by the above method strongly depends on tuning parameters λ and α . In order to decide appropriate values for them, we apply a model selection criterion. Although the cross validation is commonly used for selecting such parameters, it needs multiple computations for estimation and may often be computationally expensive. On the other hand, various criteria based on information criterion or Bayesian information criterion (BIC) are used to evaluate models from viewpoints of prediction accuracy and model selection consistency. Here we apply a BIC-type model selection criterion. Zhang et al. (2010) showed that the BIC-type criterion consistently select models when we apply the SCAD penalty (Fan and Li, 2001). The effective degrees of freedom is obtained by the trace of the (pseudo) smoother matrix. The smoother matrix of our model is obtained by calculating

$$\begin{aligned} \tilde{Z}_j \hat{b}_j &= \tilde{Z}_j R_j^{-1} C_j Q_j^T W_j^{1/2} r_{-j} \\ &= S_j r_{-j}, \end{aligned}$$

where $S_j = \tilde{Z}_j R_j^{-1} C_j Q_j^T W_j^{1/2}$ and C_j is given by

$$C_j = \begin{pmatrix} c_{j1} 1_{M_j}^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_{j(L-1)} 1_{M_j}^T \end{pmatrix},$$

$$c_{jl} = \frac{\|\hat{b}_j^*\|_2 (\|\tilde{r}_{-jl}\|_2 - n\alpha\lambda)_+}{\|\hat{b}_j^*\|_2 + n(1-\alpha)\lambda} \frac{1}{\|\tilde{r}_{-jl}\|_2}.$$

We consider S_j as a smoother matrix of our model, and therefore the effective degrees of freedom is given by $df = \sum_j \text{tr} S_j I(\|\hat{b}_j\|_2 \neq 0)$. Thus we have a model selection criterion

$$\text{BIC} = -2\ell(\hat{b}) + df \log n.$$

We choose the values of λ and α that minimize BIC and then regard the corresponding model as an optimal model.

4 Example with real data

We applied the proposed method to the analysis of yeast cell cycle gene expression data. Spellman et al. (1998) measured expression profiles over about two cell cycles for 6,178 genome-wide yeast genes using cDNA microarrays. The data contain 77 microarrays with several types of temporal synchronization: *cln3* (2 points), *clb2* (2 points), α -factor (18 points), *cdc15* (24 points), *cdc28* (17 points), and *elu* (14 points). Spellman et al. (1998) used the clustering method from the above 77 experiments to classify 800 genes into 5 groups: G1, G2/M, M/G1, S, and S/G2. Figure 1 shows examples for each type of synchronization. Araki et al. (2009a) classified genes by using the *cdc15* experiments as functional data and then used the posterior probabilities to determine the misclassified data. Here we consider if these 6 experiments affect the classification.

Since there are many missing values in the expression profiles and only 72 genes have no missing values, we excluded genes according to the following two rules: (1) Genes with at least one missing value for either *cln3* or *clb2* were excluded. (2) Those with a total of more than 10 missing values from some combination of α -factor, *cdc15*, *cdc28*, and *elu* were excluded. We can easily apply the regression model even if there are some (not excessively many) missing values by converting them into functional data. The resulting 657 genes were used for this analysis. First, except for *cln3* and *clb2*, we smoothed the time-course data to construct functions. They were expressed using basis expansions with 4 basis functions that were previously selected. The remaining variables, *cln3* and *clb2*, each of which have only 2 time points, were treated as vector data rather than functional data. We also treated the variables corresponding to the 2 time points as a group. Then

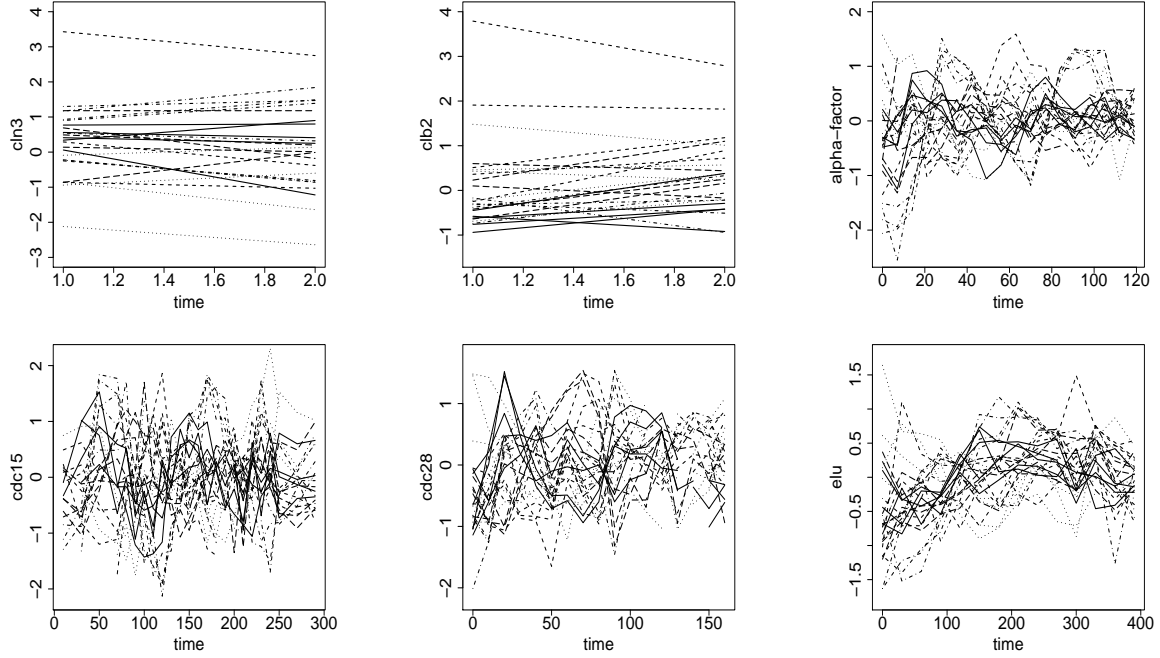


Figure 1: Yeast cell cycle gene expression profiles for each type of synchronization. Each plot consists of 5 genes from 5 classes: G1 (solid), G2/M (dashed), M/G1 (dotted), S (dot-dashed), and S/G2 (long dashed).

we constructed a functional logistic regression model as follows:

$$\log \left\{ \frac{\Pr(g_i = l | x_i)}{\Pr(g_i = L | x_i)} \right\} = \beta_{0l} + \sum_{j=1}^2 \sum_{j'=1}^2 x_{ij,j'} \beta_{j,j',l} + \sum_{j=3}^6 \int x_{ij}(t) \beta_{jl}(t) dt, \quad (9)$$

which is a special case of (1), where x_{ij} ($j = 1, \dots, 6$) correspond to gene expression profiles for cln3 , clb2 , $\alpha\text{-factor}$, cdc15 , cdc28 , and elu , respectively. The model was estimated by the penalized likelihood method with the sparse group lasso-type penalty and then the regularization parameter was selected by the BIC. We repeated this process for 50 bootstrap samples. Furthermore, we altered the class label L on the left-hand side of (9) and repeatedly estimated the model in order to investigate all the coefficients of the decision boundaries. As a result, there are totally 100 repetitions for the model for all combinations of two classes. We then investigated which variables and decision boundaries affected the classification.

Table 1 shows the numbers of selected decision boundaries for bootstrap samples. We found that many coefficients were estimated to be nonzero. However, the coefficient for the boundary between M/G1 and S/G2 was not selected at all for clb2 , and, similarly, those between M/G1 and S and S and S/G2 were rarely selected. This indicates that the variable clb2 does not affect the above classifications. On the other hand, Table 2 shows that all the variables are selected for each of the 100 repetitions in the viewpoint

Table 1: Numbers of selected decision boundaries.

	cln3	clb2	α	cdc15	cdc28	elu
G1 vs. G2/M	100	100	91	100	100	64
G1 vs. M/G1	100	60	98	70	95	95
G1 vs. S	88	64	94	99	97	100
G1 vs. S/G2	92	60	99	99	100	98
G2/M vs. M/G1	46	52	55	100	90	45
G2/M vs. S	90	55	34	99	93	52
G2/M vs. S/G2	75	51	52	100	70	45
M/G1 vs. S	63	9	55	79	71	69
M/G1 vs. S/G2	48	0	100	100	94	79
S vs. S/G2	29	2	51	55	48	51

Table 2: Numbers of selected variables.

cln3	clb2	α	cdc15	cdc28	elu
100	100	100	100	100	100

of variable selection. This result indicates that all of the variables themselves are relevant to the classification.

5 Concluding remarks

We have proposed the method for selecting both variables and decision boundaries in estimating the multiclass logistic regression model for functional data. We derived the estimation and evaluation procedures for the model with the sparse group lasso-type penalty. The model was fitted by the penalized maximum likelihood method using the blockwise coordinate descent algorithm, and then the tuning parameters involved in the model was selected by the model selection criterion. The sparse group lasso penalty is composed of two terms; the group lasso and the lasso. The former has a role of selecting variables, on the other hand, the latter selects decision boundaries. The proposed method was applied to the analysis of gene expression data, and we then investigated which types of time synchronization contributed to the classification of cell cycles.

For estimating the model with the sparse group lasso penalty, several algorithms are proposed such as Simon et al. (2013a) and Vincent and Hansen (2014). Furthermore, the alternating direction method of multipliers by Boyd et al. (2011) appears to be useful for estimating our model. We will consider the application and comparison of these methods as future works. We derived the BIC using the idea of the effective degrees of freedom,

but the BIC is originally derived from the framework of the maximum likelihood method. The derivation of the model selection criterion for our model estimated by the penalized likelihood method is also a future work.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists (B) No. 25730017 and No. 16K16020 of JSPS.

References

- Aguilera, A. and Escabias, M. (2008), “Solving Multicollinearity in Functional Multinomial Logit Models for Nominal and Ordinal Responses,” in *Functional and Operatorial Statistics*, Springer, pp. 7–13.
- Aguilera-Morillo, M. C., Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2013), “Penalized spline approaches for functional logit regression,” *Test*, 22, 251–277.
- Aneiros, G., Ferraty, F., and Vieu, P. (2011), “Variable Selection in Semi-Functional Regression Models,” in *Recent advances in functional data analysis and related topics*, Heidelberg: Springer, pp. 17–22.
- Araki, Y., Konishi, S., Kawano, S., and Matsui, H. (2009a), “Functional logistic discrimination via regularized basis expansions,” *Comm. Statist. Theory Methods*, 38, 2944–2957.
- (2009b), “Functional regression modeling via regularized Gaussian basis expansions,” *Ann. Inst. Statist. Math.*, 61, 811–833.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, 3, 1–122.
- Breheny, P. and Huang, J. (2009), “Penalized methods for bi-level variable selection,” *Stat. Interface*, 2, 369–380.
- Bühlmann, P. and van de Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Heidelberg: Springer.
- Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2007), “Functional PLS logit regression model,” *Comput. Statist. Data Anal.*, 51, 4891–4902.

- Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Amer. Statist. Assoc.*, 96, 1348–1360.
- Ferraty, F., Hall, P., and Vieu, P. (2010), “Most-predictive design points for functional data predictors,” *Biometrika*, 97, 807–824.
- Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), “Pathwise coordinate optimization,” *Ann. Appl. Statist.*, 1, 302–332.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010a), “A note on the group lasso and a sparse group lasso,” *arXiv preprint arXiv:1001.0376*.
- (2010b), “Regularization paths for generalized linear models via coordinate descent,” *J. Stat. Softw.*, 33, 1–22.
- Gertheiss, J., Maity, A., and Staicu, A.-M. (2013), “Variable selection in generalized functional linear models,” *Stat.*, 2, 86–101.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.
- Horváth, L. and Kokoszka, P. (2012), *Inference for functional data with applications*, New York: Springer.
- Konishi, S. and Kitagawa, G. (2008), *Information criteria and statistical modeling*, New York: Springer.
- Krishnapuram, B., Carin, L., Figueiredo, M., and Hartemink, A. (2005), “Sparse multinomial logistic regression: Fast algorithms and generalization bounds,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 27, 957–968.
- Matsui, H. (2014), “Variable and boundary selection for functional data via multiclass logistic regression modeling,” *Comput. Statist. Data Anal.*, 78, 176–185.
- (2015), “Sparse resregularization for bi-level variable selection,” *J. Jpn. Soc. Comp. Statist.*, 28, 83–103.
- Matsui, H. and Konishi, S. (2011), “Variable selection for functional regression models via the L1 regularization,” *Comput. Statist. Data Anal.*, 55, 3304–3310.
- McCullagh, P. and Nelder, J. A. (1989), *Generalized linear model*, London: Chapman & Hall/CRC.

- Mingotti, N., Lillo, R., and Romo, J. (2013), “Lasso variable selection in functional regression,” *Statistics and Econometrics Working Papers from Universidad Carlos III*.
- Park, M. and Hastie, T. (2007), “L1-regularization path algorithm for generalized linear models,” *J. Roy. Statist. Soc. Ser. B*, 69, 659–677.
- Ramsay, J. and Silverman, B. (2002), *Applied functional data analysis: methods and case studies*, New York: Springer.
- Simon, N., Friedman, J., and Hastie, T. (2013a), “A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression,” *arXiv preprint*, arXiv:1311.6259.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013b), “A sparse-group lasso,” *J. Comput. Graph. Statist.*, 22, 231–245.
- Simon, N. and Tibshirani, R. (2012), “Standardization and the Group Lasso Penalty,” *Statist. Sinica*, 22, 983–1001.
- Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998), “Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization,” *Mol. Biol. Cell*, 9, 3273–3297.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.
- Vincent, M. and Hansen, N. R. (2014), “Sparse group lasso and high dimensional multinomial classification,” *Comput. Statist. Data Anal.*, 71, 771–786.
- Yuan, M. and Lin, Y. (2006), “Model selection and estimation in regression with grouped variables,” *J. Roy. Statist. Soc. Ser. B*, 68, 49–67.
- Zhang, Y., Li, R., and Tsai, C. (2010), “Regularization parameter selections via generalized information criterion,” *J. Amer. Statist. Assoc.*, 105, 312–323.
- Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012), “Wavelet-based LASSO in functional linear regression,” *J. Comput. Graph. Statist.*, 21, 600–617.